Federated Registries: Issues and Approaches

ADL Implementation Fest
Aug 2009

Larry Lannom
Corporation for National Research Initiatives
http://www.cnri.reston.va.us/
http://www.handle.net/

Federation

- Federation within and across information systems is useful when
  - a set of *varying features* exists across the federates – the origin of the multiplicity
    - Includes organizational boundaries, locations, content types, etc.
  - a set of common features exists across federates – providing the value of federation
    - Shared topics, common audience, etc.
- Federation has two goals
  - Increase overall utility of the system(s) by leveraging the shared functions/features
  - Keep the distinct features of the federates, lowering barriers to collaboration and increasing overall flexibility

CORDRA

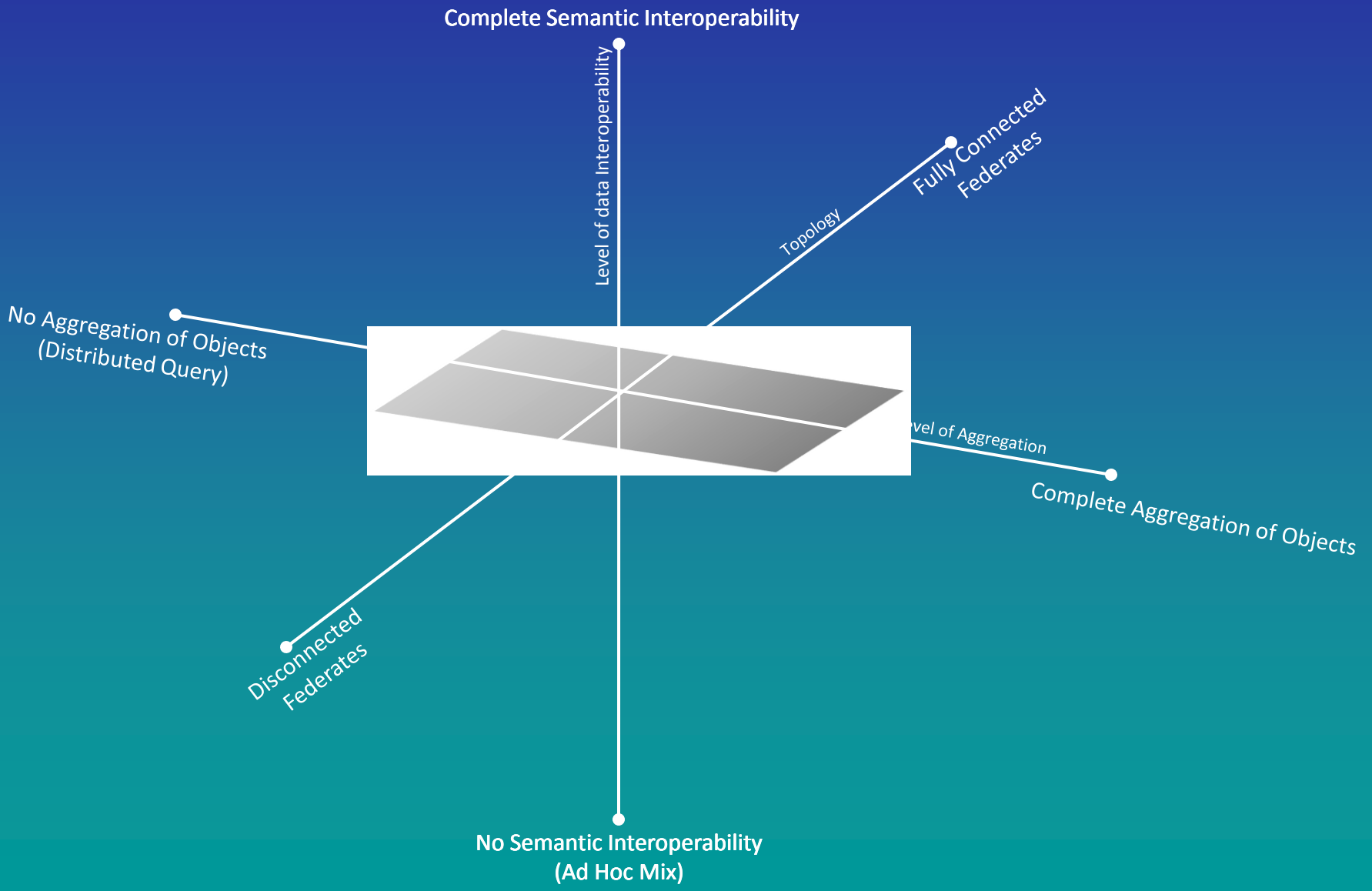
- CORDRA provides a common system/service to identify, register, and discover objects that
  - are distributed across various organizations and systems
  - are created and archived using different models/structures
  - are accessible to users through diverse mechanisms
  - but when federated form a coherent collection, i.e., the federated objects share a common information domain catering to audience interested in the collection
- Challenges in enabling CORDRA are two fold:
  - Conceptual
  - Technical

Varying features Common features

Challenges - Conceptual

- Identifying the type of aggregation:
  - Should we aggregate objects ahead of time, before any query?
  - Should we merge search responses from federates by issuing a distributed query?
  - Something in between?
- Identifying the level of semantic interoperability
  - Do we enforce complete semantic interoperability across all the data stored in the federates?
  - Do we use only the least common denominator (from a data semantics point of view) among the federates?
- Identifying the topology
  - Are all federates directly connected to each other? (fully-connected mode)
  - Is each federate connected to only its neighbor? (peer-peer mode)
- These criteria can be visualized as a *Federation Spectrum*
- Timeliness and Access Control are crosscutting issues

Role of Identifier Resolution Systems in Information Management on Networks



CORDRA Goal: One Definition

Define a framework that can be customized for federation of repositories and registries that covers all possibilities shown in the Federation Spectrum

Challenges - Technical

Depending on the criteria chosen for federation, various technical requirements arise, including:

- Design a data model to aggregate multiple metadata instances describing a single object
- Design cross-walking algorithms to translate and map heterogeneous data into a common model
- Design a query model to gather and rank search results from multiple federates
- Ensure scalability, reliability, and security without compromising performance

Digital Object Registry (basis for ADL Registry)

- Provides a *data model* to encapsulate related metadata instances together
- Enables aggregation of objects from *fully-connected* mode to *peer-peer* mode
- Uses the Handle System to uniquely and security identify objects and metadata instances across all federates

M-FASR Phase II Federation

- Use the Digital Object Registry to federate LMS repositories and ADL Registry by
  - assuming "complete semantic interoperability" of metadata
  - using aggregation mode, i.e., aggregate metadata from the ADL Registry and participating repositories at a Registry of Registries (RoR)
- Discover distributed content by searching the RoR and aggregating the discovered content into a course demonstrating the
  - Reuse of existing content to reduce cost and time
  - Repurposing of original content to meet new requirements

Future Work

- Semantic Interoperability
  - Identify a framework/mechanism for dealing with data semantics in a federation, e.g., federation specific ontologies to identify, process, and crosswalk federate specific ontologies and data
- Distributed Query and Index Aggregation
  - Look to existing IR techniques to rank search results from distributed federates
  - Aggregate indices (instead of the raw metadata) from distributed federates to optimize propagation traffic

Recent Experience

- Added dimensions to the federation challenge, beyond aggregation – topology and semantics
- Variable timeliness and depth of metadata add to the challenge
- Network and system limitations are significant in a government environment
- Meta-meta significant for aggregation – can't just combine search results – need context
- Registry of Registries now works, at least in its simplest mode

Other DO Registry/Repository Projects

- Global Environment for Network Innovations (GENI), a NSF research program, is using the DO Registry for providing clearinghouse and information management services mainly to register and discover networking resources
- DARPA Network Archive (DNA), a DARPA research program, is using the DO Repository and Registry for providing distributed storage facilities in a secure manner

Generic Registry

- We are planning on releasing the DO Registry codebase with an open source license shortly
- That codebase (aka generic registry) is a registry software package that can easily be customized to different communities and needs
- Development is completed
- Documentation efforts are ongoing